Searching for important factors in simulation models with many factors

Kleijnen, J.P.C.; Bettonvil, B.W.M.

Published in:
European Journal of Operational Research

Publication date:
1997

ELSEVIER

European Journal of Operational Research 96 (1996) 180–194

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

Searching for important factors in simulation models with many factors: Sequential bifurcation

Bert Bettonvil, Jack P.C. Kleijnen *

Department of Information Systems (BIKA) and Center for Economic Research (CentER), Tilburg University (Katholieke Universiteit Brabant), P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Received 1 October 1995; accepted 1 May 1996

Abstract

This paper deals with the problem of 'screening'; that is, how to find the important factors in simulation models that have many (for example, 300) 'factors' (also called simulation parameters or input variables). Screening assumes that only a few factors are really important (parsimony principle). This paper solves the screening problem by a novel technique called 'sequential bifurcation'. This technique is both effective and efficient; that is, it does find all important factors, yet it requires relatively few simulation runs. The technique is demonstrated through a realistic case study, concerning a complicated simulation model, called 'IMAGE'. This simulation models the greenhouse phenomenon (the worldwide increase of temperatures). This case study gives surprising results: the technique identifies some factors as being important that the ecological experts initially thought to be unimportant. Sequential bifurcation assumes that the input/output behavior of the simulation model may be approximated by a first-order polynomial (main effects), possibly augmented with interactions between factors. The technique is sequential; that is, it specifies and analyzes simulation runs, one after the other.

Keywords: Simulation; Screening; Sensitivity analysis; Design of experiments; Environment

1. Introduction

This paper presents a novel technique for *screening*, which is the process of searching for the few (say K) really important factors among the great many (say k) potentially important factors that affect a system's performance ($k \ll K$). In practice, experiments with *simulated* systems often do involve many factors. One example is the case study in this paper, which has 281 factors. This study concerns a complicated deterministic simulation model of the 'greenhouse' phenomenon (the increase in temperatures worldwide, caused by increasing quantities of carbon dioxide or CO_2 and of other gases in the atmosphere). Other examples would be stochastic simulations of queueing networks (such as logistics, computer, and telecommunication systems) with nodes that represent individual servers (machines). Large networks may have many parameters (different service rates per node), input variables (number of servers per node), and behavioral relationships (queuing discipline per node). In other words, simulation models with *several hundred factors* are common (see Morris, 1991).

* Corresponding author. Email: kleijnen@kub.nl.

PII S0377-2217(96)00156-7

The *parsimony* principle implies that the mission of science (including operations research/management science, briefly OR/MS) is to come up with a short list of the most important factors; it is unacceptable to say 'everything depends on everything else'. Additionally, screening is necessary in the *pilot phase* of complicated simulation studies. (The factors then identified as being important can be further explored in later phases; for example, the important factors might be optimized; these phases, however, are not part of this paper.)

The problem is that a simulation run may require so much *computer time* that the number of runs (say n) must be much smaller than the number of factors ($n \ll K$). The OR/MS analysts and their clients may not wish to wait until (roughly) K runs will have been generated. For example, the ecological model 'IMAGE' took 15 minutes of computer time per run, so 282 runs would have taken roughly 70 hours of computer time. (These numbers hold for the computer we had available at that time, a PC 386DX, 16 MHz. Computer speed increases steadily with technological progress, at decreasing costs. But this changing information technology raises the users' aspiration levels, so users want to explore bigger, more realistic simulation models.)

To solve the screening problem (that is, $k \ll K$, $n \ll K$), one scientific principle can be applied: *aggregation*. Indeed, aggregation has been used in the study of many large systems. For example, in economic theory and practice the production volumes of the individual companies are aggregated into the Gross National Product (GNP). More specifically, when experimenting with large systems (either real or simulated), analysts applied *group screening*: they combined individual factors into groups and experimented with these groups as if they were individual factors (details will follow). The theory on group screening goes back a long time (Jacoby and Harrison, 1962; Li, 1962; Patel, 1962; Watson, 1961). A more recent publication is Morris (1987). However, none uses 'sequential bifurcation' (SB)!

Practical applications of this group screening are rare. Our explanation is that in experiments with real systems, it is impossible to control hundreds of factors. And in simulation, most analysts seem to be unaware of group screening. Four simulation applications of group screening are summarized in Kleijnen (1987, p.327); another application is given in Cochran and Chang (1990).

This paper also uses group screening, but applies a novel design (namely, SB instead of fractional factorials such as 2^{k-p} designs) and a novel analysis (no regression analysis or analysis of variance, ANOVA), as we shall see. SB will turn out to be more efficient than competing group screening techniques. Moreover, SB has been used in practice. One application is the ecological simulation model mentioned above (in this case study, SB gave results that surprised the ecological experts; see Section 3. One other application of SB concerns a complex simulation by De Wit (1995), who studied thermal dynamics for buildings (De Wit found the 16 most important inputs among 82 factors after only 50 runs; he conformed these results by applying a different screening technique, namely Morris's (1991) 'randomized one-factor-at-a-time designs', which took as many as 328 runs).

How general are the results of these two SB applications; in other words, can SB be applied with confidence to other simulation models? Science must make *assumptions* to make progress; this also holds for screening techniques! Our experience is that in practice the simulation analysts often leave their assumptions implicit. Frequently analysts assume that they know which factors are unimportant, and they investigate only a few *intuitively* selected factors. Often they apply an inefficient and ineffective design, namely the one-factor-at-a-time design (see Van Groenendaal, 1994). All group screening techniques (including SB) assume a low-order polynomial approximation (or metamodel) for the input/output (I/O) behavior of the simulation model, and known signs or directions for the first-order or main effects. We now discuss these two assumptions in more detail.

A *polynomial approximation* implies that the underlying simulation model is treated as a *black box*. The advantage of a black box is that it can be applied to *all* types of random and deterministic simulation. The disadvantage is that it cannot exploit the special structure of the simulation model at hand (see Ho and Cao, 1991, and Rubinstein and Shapiro, 1993). The additional advantage of a *low-order* polynomial is that it is simple. Such polynomials are often used in the design of experiments (DOE) with

its concomitant ANOVA, applied to real or simulated systems. (See Kleijnen (1995) for a survey.) The classic group screening designs assume a first-order polynomial. Kleijnen (1975), however, proves that *two-factor interactions* do not bias the main effect estimators if a resolution IV design is used for the group factors. (Originally, resolution IV designs were defined, not for group factors, but for individual factors. They give estimators of all individual main effects not biased by possible two-factor interactions. The estimators for the individual two-factor interactions are biased; they estimate specific sums of these interactions.) We shall return to these designs in Section 3 (case study). In SB (as in classic group screening) a first-order polynomial requires (roughly) only half the number of runs that an approximation with two-factor interactions does. In general, however, we recommend a more cautious approach, that is, a metamodel with interactions.

Known signs of the main effects are assumed by all group screening techniques, in order to know with certainty that individual effects do *not compensate* each other within a group. In practice, the sign of a factor may be known indeed (the magnitude is unknown; hence simulation is used). For example, in queueing networks the response may be throughput per year, and the analysts may assume that higher server rates have non-negative effects on the response. In the ecological case study the experts felt confident when they had to specify the signs of the factor effects. (If the analysts feel that they do not know the signs of a few factors, then they may treat these few factors separately. Indeed, in the ecological case study there is a very small group of factors with unknown signs. These factors can be investigated in a traditional design.) We shall return to the two crucial assumptions of group screening techniques.

A characteristic of sequential procedures (such as SB) is that the analysts do not need to quantify a priori how big a factor effect should be in order to be called *important*. As simulation outputs become available, SB updates the upper limits for the factor effects; the analysts can stop the simulation experiment as soon as they find these limits sharp enough. Obviously, the analysts may make these limits depend on the system being simulated.

The main objective of this paper is to inform OR/MS analysts about a novel technique for the screening of large simulation models. We emphasize concepts; for many technical details we refer to the more than 200 pages of the doctoral dissertation (Bettonvil, 1990).

The paper is organized as follows. Section 2 explains the basics of SB, first assuming the simplest approximation, namely a first-order polynomial; next augmenting this approximation with interactions. Section 3 presents a case study, namely the ecological simulation model 'IMAGE'. Section 4 revisits the assumptions of SB. Section 5 examines the efficiency of SB. Section 6 gives conclusions, and indicates future research.

2. Basics of sequential bifurcation

Let y denote the simulation output, $s(\cdot)$ the mathematical function specified by the simulation model, and v_j the jth factor, with $j = 1,2,\ldots,K$:

$$y = s(v_1,\ldots,v_j,\ldots,v_K). \tag{1}$$

We concentrate on deterministic simulation models, which are standard in ecology, investment analysis, system dynamics, and so on.

2.1. First-order polynomial approximation and concomitant design

It is convenient to transform an original, quantitative factor v in (1) linearly into a *standardized* variable (say) x that has the values 0 and 1, where 0 (and 1 respectively) corresponds with the level that generates a low (and high) output. We concentrate on quantitative factors (for qualitative factors see Section 6, future research).

The simplest approximation of the simulation model in (1) is a first-order polynomial in the standardized variables, which has main effects (say) β_j and overall mean β_0. We define *negligible approximation errors*, as errors that are 'small' relative to the factor effects. In this section on the basics we assume that these errors are zero. Because we shall revisit the SB assumptions (see Section 4), we number all assumptions. So we formulate the following assumption.

Assumption 1. A *first-order polynomial* gives 'negligible' approximation errors over the experimental domain of the simulation model.

Hence the simulation model (1) is identical to

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_K x_K. \quad (2)$$

All group screening techniques (including SB) assume *known signs* for the main effects (see Section 1). Therefore L_j can be defined as the level of factor j that generates a low value for the output y, and H_j as the level that generates a higher value, provided this factor has any effect at all. (Consequently, if an increase of factor j reduces the output, then $L_j > H_j$; see Table 1, discussed later.) The domain of a factor is limited by the area of experimentation (also see Section 4, revisiting assumptions). This leads to the following assumption.

Assumption 2. The *direction* of the influence that a factor has on the output, *if* that factor has any effect at all, is *known*.

This assumption, together with the definition of L_j and H_j, implies that all main effects in Eq. (2) are non-negative: $\beta_j \geq 0$. In this subsection we call all positive effects *important* (they are bigger than the approximation errors, which are zero in this subsection).

Next consider the *DOE*. We introduce the symbol $y_{(j)}$ to denote the simulation output when the factors $1, \ldots, j$ are switched on and the remaining factors $(j+1, \ldots, K)$ are off. Hence, the first j original factors are at their 'high' values (see the definition of H), so the corresponding standardized variables have the value 1. The remaining $K-j$ original factors are at their 'low' values, so the corresponding standardized variables are 0. So the polynomial in Eq. (2) yields

$$y_{(j)} = \beta_0 + \beta_1 + \cdots + \beta_j, \quad j = 0,2,\ldots,K. \quad (3)$$

Consequently, the sequence $\{y_{(j)}\}$ is non-decreasing in j (this property will be discussed in Section 4, revisiting assumptions). We further introduce the symbol $\beta_{j-j'}$ to denote the sum of individual effects β_j through $\beta_{j'}$ with $j' > j$; for example, β_{1-128} denotes the sum of β_1 through β_{128} (see Fig. 1, line 1).

SB is a *sequential* procedure, that is, the selection of the next factor combinations to be simulated, depends on the outputs of previous combinations already simulated. At the start (stage #0) of the procedure, SB always observes the two 'extreme' factor combinations, namely $y_{(0)}$ (no factor high, that is, all factors low) and $y_{(K)}$ (all factors high). (In sensitivity analysis the overall mean β_0 is not of interest. Yet SB does yield an estimator of β_0, namely $y_{(0)}$; see Eq. (3) with $j = 0$. If the analysts used the polynomial approximation for prediction purposes, then β_0 would become relevant.)

We explain SB informally (for formalized procedures see Bettonvil, 1990). We use the example in Jacoby and Harrison (1962), which considers $2^7 = 128$ factors. However, only three factors are important, namely #68, #113, and #120 (obviously SB does not know this). The presence of important factors implies that, at the end of stage #0, SB gives $y_{(0)} < y_{(128)}$. Hence SB infers that the sum of all individual main effects is important: $\beta_{1-128} > 0$. SB works in such a way that any important sum of effects leads to a new observation that splits that sum into two sub-sums; also see the symbol ↓ in Fig. 1. *Bifurcation* means that the group of factors is split into two subgroups of *equal size*. This strategy assumes that K (total number of factors) is a power of two (this assumption is revisited in Section 5, on SB efficiency). In the example there are indeed 2^7 factors. Because stage #0 gives $\beta_{1-128} > 0$, SB proceeds to the next stage.

Stage #1 gives $y_{(64)}$. The SB analysis first compares $y_{(64)}$ with $y_{(0)}$, and notices that these two outputs are equal (only factors #68, #113, and #120 are important). Hence SB concludes that the first 64 individual factors are unimportant! So after only three simulation runs and based on the comparison of two runs, SB *eliminates* all factors in the first half of the total group of 128 factors. Next, SB compares $y_{(64)}$ with $y_{(128)}$, and notices that these two outputs are not equal. Hence SB concludes that the second subgroup of 64 factors is important; that is, there is at least one important factor in the second half of the group of 128 factors.

In stage #2 SB concentrates on the remaining factors (#65 through #128). That subgroup is again bifurcated, and so on. At the end, SB finds the three important factors (#68, #113, #120). In total, SB

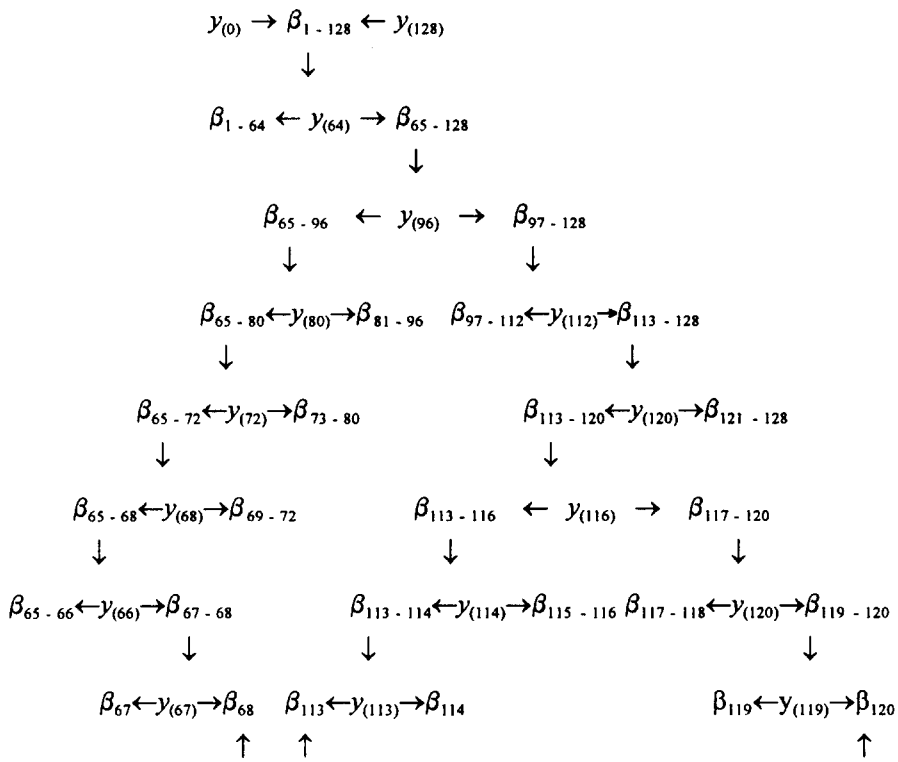

Fig. 1. Finding $k = 3$ important factors among $K = 128$ factors in Jacoby and Harrison's (1962) example.

requires only 16 observations. SB also determines the individual main effects of the important factors: see the symbol '↑' in the last line of Fig. 1.

In general, the SB *design* specifies that one new observation be generated by switching on *all* factors from #1 up to halfway the important original subgroup; for example, in stage #2 SB switches on the factors #1 through #96. The SB *analysis* compares this new observation with two old observations, to infer the two sums of the effects in the two smaller sub-subgroups; for example, in stage #2 $y_{(96)}$ splits β_{65-128} into β_{65-96} and β_{97-128}. (If there is only one important factor, then SB is simply classical binary search; see, for example, Brassard and Bradley (1988).)

2.2. Quantifying importance

In Section 2.1 SB declared a factor j important if its effect β_j was positive (not zero). Now consider the example in Fig. 2. (We found this example in the literature, so it is not meant to favor SB. Unfortunately, we do not remember which publication is the source of this example.) There are $K = 24$ factors; the sizes of their main effects are shown by the vertical bars. Before explaining the details of this figure, we mention that SB may be interpreted through the following analogy. The water level in a water basin is lowered, so that the most important (highest) rocks show up first. After the water level has been lowered for a while, the obvious conclusion is that the remaining rocks that remain below the water level, are unimportant (not high).

Stage #0 (not displayed in Fig. 2) yields $y_{(0)} = 0.0$ and $y_{(24)} = 2388.2$. Hence after two observations SB gives an upper limit for each individual effect, namely $2388.2 - 0.0 = 2388.2$. We denote the upper limit after two observations by $U(2)$. If the users were interested only in effects exceeding 2388.2, then SB could stop! Suppose, however, that the users are also interested in effects smaller than 2388.2.

If a group size is not a power of two (here: 24), then SB does not split that group in the middle (here: 12). Instead, the first subgroup gets a size equal to

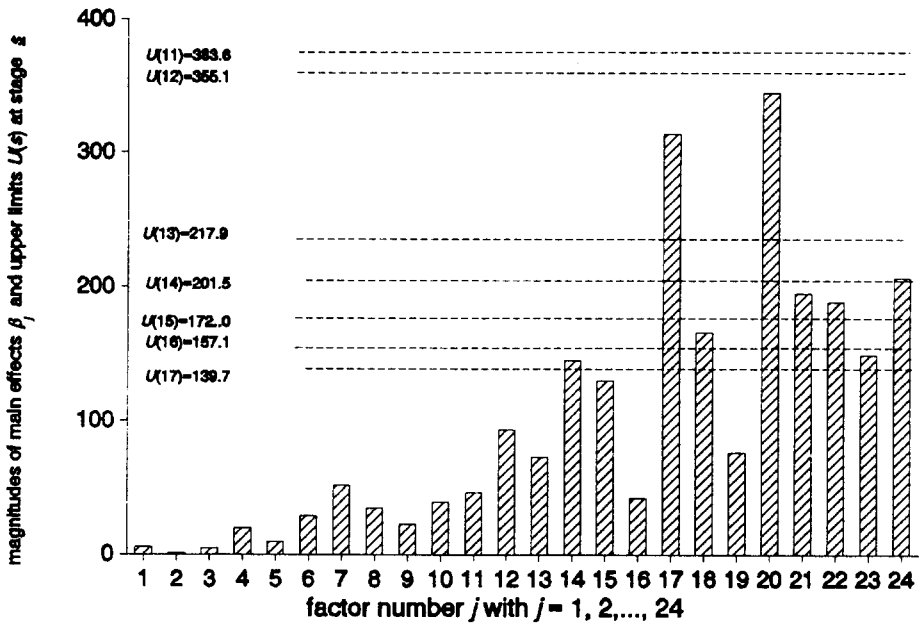

Fig. 2. Main effects β_j ($j = 1, \ldots, 24$) and upper limits $U(s)$ after stage s.

the largest possible power of two; in this case the 24 factors are split into two groups, one with 16 ($= 2^4$) factors, and the remaining group with 8 ($= 24 - 16$) factors. This type of splitting improves efficiency (see Section 5.1, also covering group − sizes not equal to a power of 2).

So in stage #1 SB takes observation #3, namely $y_{(16)}$. This observation is compared with the two preceding ones: SB computes $y_{(24)} - y_{(16)}$ and $y_{(16)} - y_{(0)}$. We concentrate on the largest difference, which is $U(3) = y_{(24)} - y_{(16)}$. At this stage this value is the most stringent upper limit available for *all* individual effects (β_1 through β_{24}). To make this limit sharper, SB takes an observation within the group that gives the sharpest limit so far: SB observes $y_{(20)}$. Again SB compares this observation with its two predecessors, which gives $y_{(24)} - y_{(20)}$ and $y_{(20)} - y_{(16)}$. Now SB considers all groups that have not been split so far, and computes the largest difference over these groups:

$$U(4) = \max\{y_{(24)} - y_{(20)}, y_{(20)} - y_{(16)}, y_{(16)} - y_{(0)}\}.$$

Obviously this new maximum does not increase, as more observations are generated.

In general, as SB proceeds through the successive stages, the groups get smaller and the upper limits decrease. The dotted horizontal lines in Fig. 2, denoted by $U(11), \ldots, U(17)$, are the upper limits after $11, \ldots, 17$ observations. These limits decrease from $U(11) = 383.6$ to $U(17) = 139.7$. After 13 observations SB reveals that the factors #17 and #20 are 'important'; that is, these two factors have effects larger than $U(13) = 217.9$, whereas the remaining factors have effects smaller than 217.9. After 17 observations SB identifies the eight most important factors (#14, #17, #18, and #20 through #24), all exceeding 139.7; the remaining factors have effects smaller than 139.7.

This example demonstrates a characteristic of sequential methods such as SB: the analysts or their clients do not need to specify a prior critical value (say) δ that must be exceeded by an effect in order to be called important. Instead, as soon as the users consider a sum of effects to be small, the investigation of this subgroup can be stopped. Their critical value is not completely subjective: δ does depend on the problem at hand (also see Section 3).

2.3. Two-factor interactions and concomitant design

By definition, interaction means that the effect of a specific factor depends on the levels of other

factors. In this section we replace Assumption 1 by the following assumption.

Assumption 1*. A first-order polynomial augmented with *cross-products* between factors gives 'negligible' approximation errors over the experimental domain of the simulation model.

Hence the simulation model in Eq. (1) is approximated, not by Eq. (2), but by

$$y = \beta_0 + \sum_{j=1}^{K} \beta_j x_j + \sum_{j=1}^{K-1} \sum_{j'=j+1}^{K} \beta_{j,j'} x_j x_{j'}. \quad (4)$$

Actually, we use a different *parametrization*, namely $z_j = -1$ iff $x_j = 0$, and $z_j = 1$ iff $x_j = 1$ (explained below Eq. (8)). Replacing x by z in Eq. (4) gives

$$y = \gamma_0 + \sum_{j=1}^{K} \gamma_j z_j + \sum_{j=1}^{K-1} \sum_{j'=j+1}^{K} \gamma_{j,j'} z_j z_{j'}, \quad (5)$$

where old and new factor effects are related as follows:

$$\gamma_{j,j'} = \tfrac{1}{4}\beta_{j,j'}, \quad j = 1,\ldots,K-1, \quad j' = j+1,\ldots,K, \quad (6)$$

$$\gamma_j = \tfrac{1}{2}\beta_j + \tfrac{1}{4}\sum_{j' \neq j} \beta_{j,j'}, \quad (7)$$

$$\gamma_0 = \beta_0 + \tfrac{1}{2}\sum_{j=1}^{K} \beta_j + \tfrac{1}{4}\sum_{j=1}^{K-1} \sum_{j'=j+1}^{K} \beta_{j,j'}. \quad (8)$$

To decide between the model in x and the one in z, we reason as follows. SB uses an experimental area that is defined by a K-dimensional rectangle. SB's DOE implies that the underlying system is simulated (observed) for combinations of *extreme* values for the simulation's input factors v_j with $j = 1,\ldots,K$ (these combinations correspond with the corners of the experimental area). To the resulting I/O data we may fit the metamodel in x or the model in z; see Eq. (4) and Eq. (5). These two models belong to the same type of metamodel (first-order polynomial plus two-factor cross-products). Their coefficients (β and γ), however, have different values; see Eqs. (6)–(8). Which coefficients represent relevant factor effects? It is nonsense to measure effects at the *origin* of the original factors v_j, which is the point $(0,\ldots,0)$. For example, when analysts investigate an anti-fever medicine, then they should experiment with (simulated) patients who have a temperature between (say) 37° Celsius and 42°, not between 0° and 100°. Further Bettonvil and Kleijnen (1990) prove that the model in x measures factor effects at a corner of the experimental area, namely, the *lower-left corner* where $x_j = 0$ for all j. Factor effects should be measured at the *center* (midpoint) of the experimental area for v_j, not at a specific corner. Hence the model in z is relevant, not that in x. Both models give essentially the same factor effects if there are no interactions: $\gamma_j = \frac{1}{2}\beta_j$; see Eq. (7) (the factor 2 is well-known in DOE). In the preceding subsection we used x because this simplified the explanation of SB.

In the first-order approximation all main effects were assumed non-negative: $\beta_j \geq 0$ (see Eq. (2) and Assumption 2: known directions). What does *known directions* of factor effects mean when there are *interactions*? The standard literature on DOE and ANOVA defines the *main effect* of factor j as the difference between:

(a) the average output when that factor is switched *on*, when averaging over all 2^{K-1} combinations of the remaining $K-1$ factors, and

(b) the average output when that factor is switched *off*, again averaging over all 2^{K-1} combinations.

It is easily verified that the metamodel with interactions in Eq. (5) implies that the main effect of factor j is given by $2\gamma_j$. We call a factor *important* if and only if its main effect is 'important'. Therefore we replace Assumption 2 by the following.

Assumption 2*. All K main effects in Eq. (5) are *non-negative*: $\gamma_j \geq 0$.

Next consider SB's *design for interactions*. In this case the number of runs is *doubled*: if in the first-order approximation SB observed $y_{(j)}$, now SB also observes $y_{-(j)}$, which denotes the output with the first j factors switched *off* and the remaining factors switched on. (Obviously, this definition implies $y_{-(0)} = y_{(K)}$ and $y_{-(K)} = y_{(0)}$.) We call $y_{-(j)}$ the *mirror* observation of $y_{(j)}$. (Doubling the number of runs in order to identify main effects in the presence of two-factor interactions resembles the foldover principle in Box and Wilson (1951, p.35).)

It is simple to prove that now the *difference* $y_{(j)} - y_{-(j)}$ is a non-decreasing function of j. So, if application of SB shows non-monotonic behavior for these differences, then the assumptions of SB are obviously violated (see Section 3 and Section 4). Moreover, estimators of the main effects that are *not biased* by any two-factor interaction are given by the following *successive differences* or *contrasts* (linear combinations of observations with sums of weights equal to zero):

$$(y_{(j)} - y_{-(j)}) - (y_{(j-1)} - y_{-(j-1)}) = 4\gamma_j. \quad (9)$$

In stage #0, SB still takes the old observations, $y_{(0)}$ and $y_{(N)}$. In view of Eq. (9), SB now computes

$$(y_{(K)} - y_{-(K)}) - (y_{(0)} - y_{-(0)}), \quad (10)$$

which equals $2y_{(K)} - 2y_{(0)}$ (because $y_{-(0)} = y_{(K)}$ and $y_{-(K)} = y_{(0)}$). Suppose the analysts find this contrast to be 'important'. Then SB proceeds to the next stage.

In stage #1, SB obtains both $y_{(K/2)}$ and its mirror observation $y_{-(K/2)}$. Because of Eq. (9), SB computes

$$(y_{(K/2)} - y_{-(K/2)}) - (y_{(0)} - y_{-(0)}) \quad (11)$$

for the first subgroup, and

$$(y_{(K)} - y_{-(K)} - (y_{(K/2)} - y_{-(K/2)}) \quad (12)$$

for the second subgroup. Iff Eq. (11) yields zero, SB eliminates the first subgroup; otherwise that group is bifurcated in the next stage. The contrast in Eq. (12) is treated analogously; and so on. Finally, SB arrives at the individual important factors. For these factors, SB computes unbiased main effect estimators; see Eq. (9).

In summary, it is easy to translate SB for a metamodel with 'main effects only' into SB for a metamodel with interactions.

3. Case study: The ecological simulation model 'IMAGE'

To demonstrate the application of SB in practice, we apply SB to a large ecological simulation model. This model has *not* been constructed for the purpose of testing SB, so this test of SB is a 'fair' test indeed. Because the ecological experts are interested in screening their model, they are prepared to give the particular information that SB needs.

The acronym *IMAGE* stands for 'Integrated Model to Assess the Greenhouse Effect'. It is a deterministic simulation model. IMAGE was developed by the Dutch 'National Institute of Public Health and Environmental Protection' (in Dutch: RIVM). The model consists of many modules or submodels (see Rotmans, 1990, for details). An essential part of IMAGE represents the worldwide circulation of CO_2 in the atmosphere, the oceans, and the terrestrial biosphere. We apply SB to this part only. Hence the output y denotes the simulated CO_2 concentration in the year 2100 (y is measured in parts per million, ppm). There are as many as $K = 281$ factors. We specify a metamodel with two-factor interactions; see Eq. (5). The ecological experts (not we) must specify the ranges of these 281 factors; see Table 1 for examples. Stage #0 gives $y_{(0)} = 988$ and $y_{(281)} = 1496$. It seems obvious that the difference between these two outputs is important; so SB proceeds to stage #1.

Because the group size is not a power of two (here: 281), SB gives the first subgroup a size of 256 ($= 2^8$) factors, and the second group a size of 25 ($281 - 256$). Stage #1 gives $y_{(256)}$ and its mirror observation $y_{-(256)}$ (see Section 2.3).

SB goes on, until 77 pairs of observations are simulated. By then, the upper limits for the main effects γ_j have decreased to 2.5 (analogous to U in Fig. 2). SB has then identified 15 factors with main effects that exceed 2.5; see Table 1. (The labels in Table 1 denote the order in which the factors are input to the simulation program; these labels do not indicate the importance of factors. Some ranges start from their high values to make the signs positive; an example is factor #243.)

So the remaining 266 ($= 281 - 15$) factors have effects smaller than 2.5. The 15 most important factors do not explain all variation in the output; the other 266 factors do contribute to the response (SB is meant for sensitivity analysis, not prediction). However, environmental policy should concentrate on the most important factors!

Can it be proved that the results of SB are *correct* for this case study? Unfortunately, this is a realistic case study, so the simulation model is so complex that nobody really knows the true factor effects!

Table 1
Important factors in IMAGE, found by SB after 2×77 runs

Label	Effect$_j$	Range		
j		L_j	H_j	Meaning
250	12.7475	0.0	0.6	shift from temperate forest to agricultural land
246	8.3725	0.9	1.4	shift from temperate forest to grassland
19	7.35	19.0	21.0	residence time in (thick) cold mixed layer
237	7.0925	0.0	0.3	shift from temperate forest to open tropical forest
243	6.8	0.5	0.0	shift from human area to temperate forest
242	5.6	0.4	0.0	shift from agricultural land to temperate forest
241	5.26	0.5	0.0	shift from grassland to temperate forest
240	5.2075	0.0	0.3	shift from open tropical forest to temperate forest
281	4.8305	0.41	0.38	biotic stimulation factor
13	4.46	0.81	0.79	rate of precipitation of carbon in the oceans
86	3.5525	0.11	0.095	fraction of charcoal formed upon burning of branches
239	3.4175	0.2	0.0	shift from closed tropical forest to temperate forest
22	3.4125	1282.23	1482.23	initial area of ecosystem 1 (tropical closed forest)
20	3.075	2.27	2.37	circulating mass flow (Gordon flow)
244	2.8625	0.2	0.0	shift from semi-desert to temperate forest

(Academic examples with known factor effects were presented in Fig. 1 and Fig. 2.) Initially (before being confronted with the SB results) the ecological experts expected eight factors to be important, namely the factors labeled #250, #246, #19, #281, #13, #86, #20 (the meanings of these factors are given in Table 1) plus a factor called DIFF (not in Table 1). In other words, seven of the experts' factors also are declared important by SB, and one of their factors is not declared important by SB. We try to verify these SB results as follows.

We perform an additional experiment with only these eight factors. For such a small number of factors it is computationally feasible to perform a *resolution IV design*: only 16 runs are needed. Table 2 denotes factor #250 as '(1)',..., #20 as '(7)' (left hand side of table). The symbol '(0)' denotes the grand mean γ_0 (right hand side of table). The symbol (1,2) denotes the interaction between the factors '(1)' and '(2)'. Actually, resolution IV implies that the estimator of (1,2) is biased by other two-factor interactions.

Table 2
Important factors in IMAGE, found by resolution IV design

Label	Effect	Label	Effect
(1) #250	13.07	(0)	1224.33
(2) #246	8.54	(1,2)	-0.05
(3) #19	7.31	(1,3)	0.10
(4) #281	5.16	(1,4)	0.09
(5) #13	4.42	(1,5)	0.08
(6) #86	3.54	(1,6)	0.07
(7) #20	3.10	(1,7)	0.07
(8) DIFF	1.98	(1,8)	0.06

Comparing Tables 1 and 2 shows that SB gives roughly the same point estimates as the resolution IV design does. SB (Table 1) does select seven factors expected to be important by the ecologists (Table 2). Table 2 implies that the factor DIFF is 'unimportant', namely smaller than 2.5 (upper limit in SB after 2×77 observations). Table 1 has fifteen 'important' factors, whereas Table 2 has fewer factors. So some of the important factors detected by SB were neglected by the ecologists. These results of the SB technique surprised the ecological experts, and provided new insight!

Note. SB can give an indication of the importance of two-factor interactions, as follows. Estimate the main effects from the original and the mirror observations (see Eq. (9)), and from the original observations only (see Eq. (3)). If both estimates are roughly equal, then most likely there are no two-factor interactions. Then SB may save runs by not generating mirror observations any longer. In the case study, both SB and the resolution IV design suggest that two-factor

interactions are unimportant. Hence, in hindsight, the number of runs in SB could have been halved, sticking to the design for a metamodel with main effects only. Moreover, it seems unlikely that DIFF is important through its interactions with other factors.

4. Assumptions revisited

SB uses two types of assumptions, labelled Assumption 1 or 1* and 2 or 2* (see Section 2). In this section we revisit these assumptions.

4.1. Assumption 1 / 1*: Low-order polynomial approximation

Assumption 1 stated that the first-order polynomial in Eq. (2) gives zero approximation errors over the experimental domain of the simulation model; assumption 1* claimed that adding two-factor interactions as in Eq. (5) gives zero approximation errors. Group-screening focuses on *overall* effects, not local (marginal) effects: what happens if a factor changes from its low level L to its high level H? We consider three cases.

(i) *Monotonic response surfaces.* Monotonicity means that the simulation output increases as an input increases. This case is handled well by SB.

(ii) *Non-monotonic response surfaces: quadratic effects.* The output might happen to be approximately the same at L_j and H_j, which are the two extreme values of factor j investigated by SB. In that case SB falsely infers that factor j is unimportant! Non-monotonicity may be quantified by quadratic effects: add $\gamma_{j,j} z_j^2$ to the metamodel in Eq. (5). The quadratic effects $\gamma_{j,j}$ can be estimated by observing output at a point halfway between the two extremes. We have not investigated this complication, but it certainly deserves future research (also see Andres and Hajas, 1993).

(iii) *Interactions only.* Consider the following academic example: the response y is determined solely by the interaction between two factors: $y = \gamma_{1,2} z_1 z_2$. Then SB fails.

The conclusion from (ii) and (iii) is that SB is *not a panacea.* But then again, no other technique is a remedy, since any OR/MS technique requires assumptions. We propose that in order to make progress, the OR/MS analysts explicitly specify an approximation to the complicated, nonlinear I/O behavior of their simulation model. Taking into account the time that is available for experiments with the simulation model, they should then select an appropriate technique.

Assumptions 1 and 1* mention the *experimental area.* That area should be selected such that the simulation model is *valid* in that area. (This area is called the 'experimental frame' in Zeigler (1976).) Indeed, in the case study we had to reduce the initial experimental area, because the two initial responses $y_{(0)}$ and $y_{(K)}$ (which correspond with all inputs 'off' and 'on' respectively) were unrealistically low and high respectively (the ecologists rejected these two response values immediately). Note that the importance of the factors may change as the experimental area changes.

Assumptions 1 and 1* also mention *zero approximation errors*, which means perfect fit. An alternative assumption represents the approximation errors (say) ε by *white noise*: errors are normally and independently distributed (n.i.d.) with zero expectation and constant variance (say) σ^2. Moreover, these errors are additive. Such a statistical model is standard in DOE and ANOVA. Indeed this assumption has been used by Kleijnen et al. (1992), applying standard DOE and ANOVA for sensitivity analysis of a smaller part of the IMAGE model. For SB with white noise Bettonvil (1990, pp.49–142) gives three statistical techniques. These techniques work only if the signal/noise ratio is high: $\gamma_j/\sigma > 6$. For such high ratios, however, SB may as well ignore noise, since there are so many factors in screening that only those factors with high signal/noise ratios should be detected.

SB might use a *covariance stationary process* instead of white noise, to model the approximation errors. Such a process is used to model the systematic effects of the inputs, not the noise, by Sacks et al. (1989). Their model also is used by Welch et al. (1992). Their model, however, is more complicated than our's. Moreover, it is applied to situations with only twenty factors. If the simulation model itself is stochastic, then the intrinsic noise may exceed the fitting errors. We have not investigated these complications.

4.2. Assumption 2 / 2*: Known signs

Assumption 2 stated that all K main effects in the first-order approximation are non-negative ($\beta_j \geq 0$); Assumption 2* stated that all main effects in the metamodel with two-factor interactions are non-negative ($\gamma_j \geq 0$). So these assumptions postulate *known signs* of the main effects.

Section 1 (the Introduction) has already pointed out that in practice the analysts often do know the signs of (most) factors, as was demonstrated by the ecological case study and the queueing examples. When in practice the analysts feel that they really do not know whether a particular factor meets this assumption, then they can treat that factor outside SB. An example was provided by the case study (Section 3). When the analysts falsely think that they do know the sign of a factor, then the simulation response might show so: the response decreases when this factor is switched on, provided its negative effect exceeds the sum of the positive effects of the other factors within its group (see Eq. (3)). In the case study such a decrease happened twice; however, the response decreased so little that we decided to ignore it. In general, a factor may have a negative sign, but its absolute magnitude may be negligible, practically speaking.

5. Efficiency of SB

The preceding sections covered the effectiveness of SB: can SB find all the important factors? This section is on efficiency, measured by the number of *simulation runs*.

Note. Several other measures may be distinguished. An example is the precision of the main effect estimators. Assumption 1*, however, states that the approximation errors are negligible. Hence the metamodel in Eq. (5) has no error term. Consequently, the main effects estimated by Eq. (9) are assumed to be 100% precise. Other group-screening techniques also assume negligible error terms. An efficiency measure different from statistical efficiency is 'handling efficiency': sequential procedures require that the analysts switch back and forth between the simulation program on one hand, and the analysis and design program on the other hand. In that respect, two-stage procedures are more efficient than purely sequential procedures. The development of parallel computer architectures may further complicate the implementation of sequential procedures (also see Kleijnen and Annink, 1992).

This section first covers bifurcation and clustering of important factors. Next this section compares SB with other group-screening techniques.

5.1. Bifurcation and clustering

Section 2.1 assumed that the total number of individual factors is a power of two: $K = 2^m$ with an appropriate integer m. Even if this assumption does not hold, it might seem obvious to split the factors (nearly) equally. For example, a group of six factors might be split into two groups of three each (in the next stage a group of three is split into groups of two and one). This example may be interpreted as follows: add *dummy factors* at the positions four and eight. However, it turns out that it is more efficient to split the group of six factors into a group of four and a group of two (see Bettonvil, 1990, pp.40–43). This may be interpreted as: add all dummy factors at the end of the original group. In other words, the *clustering* of (un)important factors improves efficiency!

In general, if the size of a group is not equal to a power of two, then its first subgroup should have a size equal to the largest possible power of two. For example, in the ecological case study the 281 factors were split into two groups, one with 256 ($= 2^8$) factors, and the remaining group with 25 factors.

A consequence is that SB is most efficient if the original factors are labelled from 1 through K in *increasing* order of importance, so that after stage #1 the important factors are clustered in the second group. In this way SB can take advantage of prior knowledge about the sizes (besides the signs) of factor effects. In practice, the experts' prior guess is not exact: simulation is done, because this knowledge is weak! Fortunately, a wrong guess affects only SB's efficiency, not its effectiveness (also see the next subsection, Section 5.2).

5.2. Comparison of SB with other group-screening techniques

To compare SB and its competitors, we consider three situations (scenarios): no clustering (worst

Table 3
Worst case: *maxn*; $K = 1024$ factors

Technique	Number of important factors k								
	0	1	2	3	4	5	6	7	8
Two-stage group screening	4	68	96	116	136	148	160	172	188
Multi-stage group screening [a]	2	20	35	49	62	74	85	96	107
Morris's group screening	2	12	33	53	73	80	89	100	113
Jacoby and Harrison's SB	3	21	39	55	71	85	99	113	127
SB	2	12	21	29	37	44	51	58	65

[a] Mean case for multi-stage group screening.

case), perfect clustering (best case), and the 'expected' situation (corresponding with some probability (say) p of important factors being scattered, discussed below). Pessimists choose the screening technique that scores best under the worst-case scenario; optimists base their choice on the best-case results; rationalists are guided by the expected (mean) performance.

Our computations use the two fundamental assumptions common to all group-screening techniques, namely Assumptions 1 and 2. For the worst-case scenario Table 3 gives *maxn*, the *maximum* number of runs needed to find the k important factors among the $K = 2^m$ factors. In that table *two-stage group screening* denotes the technique introduced by Watson (1961). Its performance was evaluated by Mauro (1984). (Two-stage group-screening designs are more efficient than random designs; see Mauro and Burns (1984).) *Multi-stage group screening* is a generalization of two-stage group screening, introduced by Patel (1962) and Li (1962). Both two-stage and multi-stage group screening assume that each factor has a prior probability (say) p of being important, which yields an optimal group size and, in the case of multi-stage group screening, an optimal number of stages. The table gives results for the 'optimal' guess of p (the guess that minimizes the number of observations), namely k/K. Moreover, for multi-stage group screening we use the formula for $E(n)$, the expected number of runs, which is smaller than *maxn*, the worst case number. To the best of our knowledge, their is no formula to compute the maximum (or the minimum) number of observations for multi-stage group screening. *Morris's group screening* is a variant of two-stage group devised by Morris (1987). Its number of runs also depends on p. For Morris's technique Bettonvil (1990) derives the formula for *maxn* (because it is a complicated formula that requires many details on Morris's technique, we do not display that formula). *Jacoby and Harrison's SB* uses two runs where SB uses a single one: Jacoby and Harrison (1962) use a less efficient design at each stage. For SB, *maxn* is given by

$$maxn = 1 + k\left[\log_2(2K/k)\right], \tag{13}$$

provided the number of important factors is a power of 2; otherwise this equation gives an approximation. For example, Fig. 1 ($K = 128$ and $k = 3$) gives $maxn = 20.2$, whereas the actual number of runs is

Table 4
Best case: *minn*; $K = 1024$ factors

Technique	Number of important factors k								
	0	1	2	3	4	5	6	7	8
Two-stage group screening	4	68	68	68	68	68	68	68	68
Morris's group screening	2	12	12	13	13	16	16	16	16
Jacoby and Harrison's SB	3	21	21	23	23	27	27	29	29
SB	2	12	12	13	13	15	15	16	16

16; the actual number is smaller, because the important factors show some clustering. Table 3 shows that, even though the assumptions used for that table favor the other techniques, SB has the best worst-case performance. This table gives numerical data for $K = 1024$, but for other values of K this conclusion also holds (see Bettonvil, 1990, pp.25–32).

Note. Search linear models were introduced by Srivastava (1975), assuming that the number of important factors is small but *known*. He gives specific designs only for k equal to one or two. Ghosh (1979) generalizes this approach to multi-stage designs in such a way that his technique resembles SB. His number of runs equals that of SB. Because his technique uses more restrictive assumptions than SB does, we will not further consider this technique.

Next Table 4 gives results similar to the preceding table, but now for the *best-case* scenario: *minn*. This table shows that also under this criterion SB dominates.

Finally, Table 5 gives the *expected* number of runs, $E(n)$. Morris (1987) gives this number for two-stage group screening and for his own technique. Bettonvil (1990) derives, for SB with p (prior probability) and $K = 2^m$,

$$E(n) = 1 + K - \sum_{j=1}^{m} 2^{m-j}(1-p)^{2^j}. \qquad (14)$$

This table shows that SB is also more efficient under this criterion. (Note that $p = 0.1$ is a very high value for screening.)

6. Conclusions and future research

6.1. Conclusions

The problem discussed in this paper is *screening*; that is, simulation models may have many factors (namely, K), whereas only a few factors (say, k) are really important; computer time permits only a number of simulation runs (denoted by n) *smaller* than the total number of factors. In summary, screening means: $k \ll K$, $n \ll K$.

This problem may be solved by sequential bifurcation (SB). This technique uses a design and an analysis that requires *fewer simulation runs* than other group-screening techniques (efficiency), and still finds the important factors (effectiveness).

All group-screening techniques (including SB) use *two basic assumptions*: (i) the simulation model can be approximated by a low-order polynomial metamodel, and (ii) the signs of the main effects are known.

These assumptions may be reasonable. Low-order polynomials are standard in design of experiments (DOE). SB can handle two-factor interactions easily: SB adds 'mirror' observations (foldover principle). Known signs occur in many models; an example is the ecological model IMAGE. Moreover, any other OR/MS technique also requires assumptions, which may be questionable in a particular case study. So there is no panacea.

Because SB is sequential, it can be applied without a priori quantifying what is meant by *important* main effects. As simulation outputs become avail-

Table 5
Mean case: $E(n)$; $K = 1024$ factors

Technique	Prior estimate p of fraction of important factors			
	0.0001	0.001	0.01	0.1
Two-stage group screening	21.9	64.3	198.2	521.1
Multi-stage group screening	3.6	20.2	129.2	641.9
Morris's group screening	10.3	26.4	130.3	521.1
Jacoby and Harrison's SB	4.8	19.9	138.0	745.4
SB	3.0	11.4	70.5	374.2

able, SB updates its upper limits for the main effects. The analysts can stop simulating, as soon as they judge these limits to be sharp enough, which depends on the system being simulated.

SB was applied to a complicated simulation model, namely IMAGE. This *case study* yielded results that surprised the ecological experts! The SB results were confirmed by a classical resolution IV experiment.

6.2. Future research

This paper mentioned the following issues that need further research.

(i) *Qualitative factors* (for example, priority rule) may be represented by the standardized variables z, provided the factors have only two levels each (say, FIFO versus SPT). Otherwise, the analysts may consider only those two levels that generate the minimum and the maximum outputs respectively, provided they know which two levels give extreme values.

(ii) *Quadratic effects* $\gamma_{j,j}$ create problems that are not solved by SB in its current form.

(iii) Metamodels with *interactions only* make SB fail.

(iv) SB might use a *covariance stationary process* (instead of white noise) to model the approximation errors.

(v) The simulation model may be *stochastic* (non-deterministic). Then the intrinsic noise may exceed the polynomial fitting errors.

(vi) More *expertise* should be accumulated by the application of SB and competing techniques to practical and academic simulation models that require screening.

Acknowledgements

The authors thank the two anonymous referees for their comments, which helped to improve this article.

References

Andres, T.H., and Hajas, W.C. (1993), "Using iterated fractional factorial design to screen parameters in sensitivity analysis of a probabilistic risk assessment model", presented at the Joint International Conference on Mathematical Methods and Supercomputing in Nuclear Applications.

Bettonvil, B. (1990), *Detection of Important Factors by Sequential Bifurcation*, Tilburg University Press, Tilburg.

Bettonvil, B., and Kleijnen, J.P.C. (1990), "Measurement scales and resolution IV designs", *American Journal of Mathematical and Management Sciences* 10, 309–322.

Box, G.E.P., and Wilson, K.B. (1951), "On the experimental attainment of optimum conditions", *Journal of the Royal Statistical Society. Series B* 13, 1–38.

Brassard, G., and Bradley, P. (1988), *Algorithmics: Theory and Practice*, Prentice-Hall, Englewood Cliffs, NJ.

Cochran, J.K., and Chang, J. (1990), "Optimization of multivariate simulation output models using a group screening method", *Computers and Industrial Engineering* 18/1, 95–103.

De Wit, M.S. (1995), "Uncertainty analysis in building thermal modelling", Delft Technical University, Delft, Netherlands; also in: *Proceedings SAMO95*, Theory and Applications of Sensitivity Analysis of Model Output in Computer Simulation, 25–27 September 1995, Belgirate, Italy, International Symposium organized by European Commission DGXII/F, RTD Action Energy and Joint Research Centre Environment Institute.

Ghosh, S. (1979), "On single and multistage factor screening procedures", *Journal of Combinatorics, Information and System Sciences* 4, 275–284.

Ho, Y., and Cao, X. (1991), *Perturbation Analysis of Discrete Event Systems*, Kluwer, Dordrecht.

Jacoby, J.E., and Harrison, S. (1962), "Multi-variable experimentation and simulation models", *Naval Research Logistics Quarterly* 9, 121–136.

Kleijnen, J.P.C. (1975), "Screening designs for poly-factor experimentation", *Technometrics* 17, 487–493.

Kleijnen, J.P.C. (1987), *Statistical Tools for Simulation Practitioners*, Marcel Dekker, New York.

Kleijnen, J.P.C. (1995), "Sensitivity analysis and optimization in simulation: Design of experiments and case studies", in: C. Alexopoulos, K. Kang, W.R. Lilegdon and D. Goldsman (eds.), *Proceedings of the 1995 Winter Simulation Conference*.

Kleijnen, J.P.C., and Annink, B. (1992), "Vector computers, Monte Carlo simulation, and regression analysis: An introduction", *Management Science* 38, 170–181.

Kleijnen, J.P.C., van Ham, G., and Rotmans, J. (1992), "Techniques for sensitivity analysis of simulation models: A case study of the CO_2 greenhouse effect", *Simulation* 58, 410–417.

Li, C.H. (1962), "A sequential method for screening experimental variables", *Journal of the American Statistical Association* 57, 455–477.

Mauro, C.A. (1984), "On the performance of two-stage group screening experiments", *Technometrics* 26, 255–264.

Mauro, C.A., and Burns, K.C. (1984), "A comparison of random balance and two-stage group screening designs: A case study", *Communications in Statistics. Theory and Methods* 13, 2625–2647.

Morris, M.D. (1987), "Two-stage factor screening procedures

using multiple grouping assignments", *Communications in Statistics. Theory and Methods* 16, 3051–3067.

Morris, M.D. (1991), "Factorial plans for preliminary computational experiments", *Technometrics* 33, 161–174.

Patel, M.S. (1962), "Group screening with more than two stages", *Technometrics* 4, 209–217.

Rotmans, J. (1990), *IMAGE: An Integrated Model to Assess the Greenhouse Effect*, Kluwer Academic Publishers, Dordrecht.

Rubinstein, R.Y., and Shapiro, A. (1993), *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization via the Score Function Method*, Wiley, New York.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989), "Design and analysis of computer experiments" (including Comments), *Statistical Science* 4, 409–435.

Srivastava, J.N. (1975), "Designs for searching nonnegligible effects", in: J.N. Srivastava (ed.), *A Survey of Statistical Design and Linear Models*, North-Holland, Amsterdam, 507–519.

Van Groenendaal, W. (1994), *Investment Analysis and Decision Support for Gas Transmission on Java*, Tilburg University Press, Tilburg.

Watson, G.S. (1961), "A study of the group screening method", *Technometrics* 3, 371–388.

Welch, J.W., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., and Morris, M.D. (1992), "Screening, predicting, and computer experiments", *Technometrics* 34, 15–25.

Zeigler, B. (1976), *Theory of Modelling and Simulation*, Wiley Interscience, New York.